Networking Phylogeny for Indo-European and Austronesian Languages

Philippe Blanchard[1], Filippo Petroni[2], Maurizio Serva[3] & Dimitri Volchenkov[4]

[1]*Bielefeld-Bonn Stochastic Research Center, Universität Bielefeld,Universitätsstr. 25, D-33615 Bielefeld, Germany.*

[2] *DIMADEFAS, Facoltà di Economia, Università di Roma "La Sapienza", Via del Castro Laurenziano 9, 00161 Roma, Italy.*

[3]*Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy.*

[4]*Center of Excellence Cognitive Interaction Technology, Universität Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany.*

Harnessing cognitive abilities of many individuals, a language evolves upon their mutual interactions establishing a persistent social environment to which language is closely attuned. Human history is encoded in the rich sets of linguistic data by means of symmetry patterns that are not always feasibly represented by trees. Here we use the methods developed in the study of complex networks to decipher accurately symmetry records on the language phylogeny of the Indo-European and the Austronesian language families, considering, in both cases, the samples of fifty different languages. In particular, we support the Anatolian theory of Indo-European origin and the 'express train' model of Austronesian expansion from South-East Asia, with an essential role for the Batanes islands located between the Philippines and Taiwan.

Changes in languages go on constantly affecting words through various borrowings and innovations[1]. Although tree diagrams are ubiquitous in language phylogenies, they obviously fail to reveal full complexity of language affinity; not least because of the

relation of ancestry basic for a tree structure is not always clear in cases of extreme historical language contacts[2]. Many evolutionary trees conflict with each other and with the traditionally accepted family arborescence and the languages known as isolates cannot be reliably classified into any branch with other living languages thus requiring that network models be used, instead of trees.

A network is any method of sharing information between individual units that are tied by one or more specific types of interdependency. They can often be abstracted as graphs that seem to be the natural mathematical tool for facilitating the analysis. In comparison to hierarchic tree models, the concept of proximity that formalizes the intuitive idea of closeness between two languages in a phylogenetic network implies that they share not only a recent common ancestor but the pattern of relationships with other languages in the family. Networks have already appeared in studies of the language phylogeny as a number of additional non-tree edges representing contacts between the individual language groups modelled by trees[3,4]. Here we do not follow this line, but construct a network directly from a dataset bypassing any tree-stage and analyse it using spectral methods derived in the study of complex networks[5]. Unlike most other approaches to language classification[2,4,6,7,8,9,10,11], our method based on reconstruction of symmetries of a language family does not require any previous knowledge of the language origin since it automates both the discernment on language similitude and the sagacity on the language phylogeny, with the lack of uncertainty in the estimation of tree topology, branch length estimation, and congruence between independent lines of evidence.

The key point of the approach is the definition of lexical distances among pairs of languages by a renormalized edit distance averaged over Swadesh's vocabulary of

200 meanings[9,10,11]. As a result, we have obtained two symmetric positive-definite matrices with the lexical distances between the languages in the samples of 50 major Indo-European (IE) languages and 50 Austronesian languages (AU). Then we investigated the symmetry patterns of both matrices by means of a linear transformation which can be interpreted as a random walk defined on the weighted undirected graphs determined by the matrices of lexical distances. The complete set of symmetries inherent to each language family is then encoded by a number of centred ellipsoids in Euclidean space spanned by the diffusion eigenmodes wherein languages are represented by points. The language phylogenies are revealed by geometrical proximity of the projections of language points onto the principal symmetry axes. Here, we identify two general mechanisms that explain the evolutionary dynamics of languages within a language group based on the concept of language family symmetry, a congenital birthmark that acquires meanings through a persistent process of coherent borrowings and innovations.

The Indo-European Superfamily of Languages

Examining the three major symmetries of the IE language family, we have used the relevant principal axes as a Cartesian coordinate system in 3D space, in order to obtain the representation shown in Fig.1a. In three dimensions, the IE family forms a caltrop-like structure made up of four well-separated spines representing four biggest traditional IE language groups: Romance, Germanic, Balto-Slavic, and Indo-Iranian. These groups are monophyletic and supported by the sharply localized distributions of the azimuthal and zenith angles over the languages shown in Fig.1b and Fig.1c respectively. Interestingly, the Greek, Romance, Celtic, and Germanic languages form a supergroup characterized by approximately the same azimuthal angle thus belonging to one plane (see Fig.1b), while the Indo-Iranian, Balto-Slavic, Armenian, and Albanian languages form another supergroup, with respect to the zenith angle (see Fig.1c)

attesting the Centum-Satem isogloss in the IE language family related to the evolution of the three dorsal consonant rows reconstructed for the Proto-Indo-European language[12]. The projections of Albanian, Greek, and Armenian languages onto the principal symmetry axes are rather small as they occupy the centre of the diagram in Fig.1a reflecting that their relations with other languages are not compatible with the three leading symmetries of the entire family. Being eloquently different from others, these languages can be resolved with the use of some minor symmetry. Remarkably, the Greek and Armenian languages always remain proximate confirming Greeks' belief that their ancestors had come from western Asia[13]. From time of H. Schliemann discovered Troy, four-footed symbols ubiquitous for notably all IE cultures as early as from the 5th millennium BC[14,15] were thought as pointing to their common progenitors[16]. Among the other interpretations[17], perhaps they would attest to the introspection of early Indo-Europeans on their swerving treads beyond the frontiers of the known world. In modern times, the emblem has become stigmatized in the western world and replaced by another four-footed sign, the peace symbol, coined during the campaign for nuclear disarmament in Britain, which represents the original caltrop figure better indeed.

Many language groups in the IE family had originated after the decline and fragmentation of territorially-extreme polities and in the course of migrations when dialects diverged within each local area and eventually evolved into individual languages. Geometric representations of language phylogenies can be conceived within the framework of various models that infer on the evolutionary dynamics of languages. In our three dimensional model, a complex nexus of processes beyond the emergence and diversification of languages within the group is described by the only degree of freedom, along the radial coordinate from the centre of the IE caltrop. It is worth mentioning that the distributions of languages along the radial coordinate show good agreement with univariate normality as seen from the normal probability plots shown in Fig.2a-d, for each of the major language groups. The data points were ranked and then

plotted against their expected values under normality, so that departures from linearity signify departures from normality. While the mechanisms underlying evolution of languages are very complex, the assumption of normality can be justified by taking on that many small, independent effects are additively contributing to the process. The univariant normal distribution is closely related to the time evolution of a mass-density function under homogeneous diffusion in one dimension, in which the mean value μ is interpreted as the coordinate of a point where all mass was initially concentrated, and variance grows linearly with time $\sigma^2 \sim t$. For language groups, variance grows at a very slow pace of a millionth per year giving some indication of their ages. The timing age, the ratio of the group age assessed in accordance with historically attested events[18,19] to its variance, $t/\sigma^2 = (1.367 \pm 0.002) \cdot 10^6$ has been evaluated on the bases of the last Celtic migration (to the Balkans and Asia Minor) (300 BC), the division of the Roman Empire (500 AD), the migration of German tribes to the Danube River (100 AD), and the establishment of the Avars Khaganate (590 AD) overspreading Slavic people who did the bulk of the fighting across Europe. The timing age deduced from the well-attested events allows age estimation for those groups branched off during poorly documented periods in history. In particular, the break-up of the Proto-Indo-Iranian language is estimated to happen before 2,400 BC, in a good agreement with the migration dates from the early Andronovo archaeological horizon[20]. The Balto-Slavic dialect continuum could exist sometimes before 1,400 BC supporting the recent glottochronological estimates[21] well agreed with the archaeological dating of Trziniec-Komarov culture, localized from Silesia to Central Ukraine. The group of Indo-Aryan languages had been branched off before 400 BC, probably as a result of Aryan migration across India to Ceylon, as early as in 483 BC[22]. Eventually, the Iranian languages began to break off and evolve separately as the various Iranian tribes migrated and settled in vast areas of south-eastern Europe, the Iranian plateau, and Central Asia before 400 BC, shortly after the end of Greco-Persian wars[23].

It is a subtle problem to trace back the diverging pathways of language evolution to a convergence in the IE protolanguage since symmetry of the modern languages mismatches that in ancient time. The major IE language groups have to be reexamined in order to ascertain the locations of the individual protolanguages as if they were extant. We have used variances determined from the studied samples of languages as estimators for the variance values of the entire groups and targeted the five protolanguages with the 95% confidence level (see Fig.2e). The centre point is then naturally interpreted as the expected location of the Proto-Indo-European language in space of symmetries inherent to the modern IE languages. Comparing the goodness of fit of the scaled distances from the protolanguages to the centre point to their expected values under the chi-square distribution with three degrees of freedom, we tested the locations of protolanguages for three-variant normality. As with normal probability plots, departures from three-variant normality are indicated by departures from linearity (see Fig.2f). Supposing that the underlying population of parent languages is subjected to multivariant normality, we conclude that the determinant of the sample variance-covariance matrix has to grow linearly with time. The use of the previously determined timing age then dates the initial break-up of the IE protolanguage back to 7,400 BC, in agreement with the Anatolian hypothesis of Indo-European origin[2,10,13,18,20].

In Search of Polynesian Origins by Language Symmetries

The colonization of the Pacific Islands is still the recalcitrant problem in the history of human migrations, despite many explanatory models based on linguistic, genetic, and archaeological evidences have been proposed in so far. The origins, relationships, and migration chronology of Austronesian settlers have constituted the sustainable interest and continuing controversy for decades. The symmetry probe of the 50 AU languages uncovers immediately the both Formosan (F) and Malayo-Polynesian (MP) branches of the language family (Fig.3a). The distribution of azimuthal angles

(Fig.3b) identifies them as two monophyletic jets of languages that cast along either axis spanning the entire family plane. The clear geographic patterning is perhaps the most remarkable aspect of the geometric representation. It is also worth mentioning that the language grouping recovered by language symmetry profoundly reflects historical relationships. For instance, the Malagasy language spoken in Madagascar casts in the same mould as the Maanyan language spoken by the Dayak tribe dwelling in forests of Southern Borneo and the Batak Toba language of North Sumatra spoken mostly west of Lake Toba. Despite Malagasy shares much of its basic vocabulary with the Maanyan language[24], outstandingly many manifestations of Malagasy culture cannot be linked up with the culture of Dayak people: the Malagasy migration to East Africa presupposes highly developed construction and navigation skills with the use of out-rigger canoes typical of many Indonesian tribes which the Dayak people however do not have, also some of the Malagasy cultivations and crop species (such as wet rice) cannot be found among forest inhabitants. In contrast, some funeral rites (such as the second burial, *famadihana*) widely accepted by the Malagasy culture are similar essentially to those of Dayak people. A possible explanation is that population of the Dayak origin were brought to Madagascar as slaves by Malay seafarers and unlikely realized the spectacular trip across the Indian Ocean. As the Dayak speakers formed the majority in the initial settler group, in agreement with the genetic parental lineages found in Madagascar[25], their language could have constituted the core element of what later became Malagasy, while the language of the Malay dominators was almost suppressed, albeit its contribution is still recovered by the symmetry exploration.

The AU language family forks at the northernmost tip of the Philippines, the Batanes Islands. On the distribution of azimuthal angles shown in Fig.3b, the Itbayaten language representing them in the studied sample is located pretty close to the azimuth bridging over the separating individual language family branches. By the way, the MP-offset descends from the northern Philippines (the northern Luzon Island) and springs forth

eastward through the Malay Archipelago across Melanesia culminating in Polynesia (Fig.3d); pretty in accordance with the famous 'express train' model of migrations peopled the Pacific[26]. In its turn, the F-branch embarks on the southwest coast of Taiwan and finds its way to the northern Syueshan Mountains inhabited by Atayal people that compose many ethnic groups with different languages, diverse customs, and multiple identities. Evidently, the both offshoots derived their ancestry in Southeast Asia as strengthened by multiple archaeological records, but then evolved mostly independently from each other, on evidence of the Y-chromosome haplotype spread over Taiwanese and Polynesian populations[27]. The Bayesian methods for the language phylogeny trees also evinced the earliest separation of these two branches of the AU language family. However, in the recent pulse-pause scenario, the Taiwanese origin of the entire AU family was suggested because of the "considerable diversity of Formosan languages". It is important to note that diversity itself is by no means a reliable estimate provided symmetry is downplayed (e.g., in spite of the greatest diversity, the Indo-Iranian language group is not an origin of the entire IE-Superfamily).

The distribution of languages spoken within Maritime Southeast Asia, Melanesia, Western Polynesia and of the Paiwan language group in Taiwan over the distances from the centre of the diagram representing the AU language family in Fig.3a conforms perfectly to univariate normality suggesting that an interaction sphere had existed encompassing the whole region, from the Philippines and Southern Indonesia through the Solomon Islands to Western Polynesia, where ideas and cultural traits were shared and spread as attested by trade[28,29] and translocations of farm animals[30,31] among shoreline communities. Although the lack of documented historical events makes the use of the developed dating method difficult, we may suggest that variance evaluated over Swadesh's vocabulary forges ahead approximately at the same pace uniformly for all human societies involved in trading and exchange, as presumed in traditional glottochronology. Then, the timing age deduced from the previous chronological

estimates for the IE language family returns 550 AD if applied to the AU languages as the likely break-up date of the AU dialect continuum, pretty well before 600-1,200 AD while descendents from Melanesia settled in the distant apices of the Polynesian triangle as evidenced by archaeological records[32,33,34]. The distributions of languages spoken in the islands of East Polynesia and of the Atayal language groups in Taiwan over the radial coordinate from the centre of the diagram break from normality, so that the general diffusive mechanism used previously for either of the chronological estimates is inapplicable to them. To all purposes, the evolution of these extreme language subgroups cannot be viewed as driven by independent, petty events.

Although the languages spoken in Remote Oceania clearly fit symmetry of the entire MP-branch, they seem to evolve without extensive contacts with Melanesian populations, perhaps because of a rapid movement of the ancestors of the Polynesians from South-East Asia as suggested by the 'express train' model consistent with the multiple evidences on comparatively reduced genetic variations among human groups in Remote Oceania[35,36,37]. In order to obtain reasonable chronological estimates, an alternative mechanism on evolutionary dynamics of the extreme language subgroups in symmetry space of the AU language family should be reckoned with. The simplest 'adiabatic' model entails that no words had been transferred to or from the languages riding the express train to Polynesia, so that the lexical distance among words of the most distanced languages tends to increase primarily due to random permutations, deletions or substitutions of phonemes in the words of their ancestor language. The radial coordinates of the languages at the distant margins of the family diagram shown in Fig.3a may be deduced as evolving in accordance with the simple differential equation $\dot{r} = -\alpha r$ characterised by some constant $\alpha > 0$ quantifying the rate of random phonetic changes. The suggested model of language evolution is conceived by that it is very difficult to establish whether a word has changed because of many phonemes have been replaced consequently, or the whole word has been substituted at once, without

precise historical knowledge of the languages attested relatively recently. With this choice, word substitution is statistically equivalent to the replacement of all characters in the word, and then the distance *r* cannot grow anymore. The relative dates can be derived basing on the assumption of independent random changes of phonemes in words of the ancestral language by $t_1 - t_2 = -\alpha^{-1} \ln(r_1/r_2)$ where $r_2 > r_1$ are the radial coordinates of the languages from the centre of the sample diagram (Fig.3a). Tahiti located in the archipelago of Society Islands is the farmost point in the geometric representation of the Austronesian family and the foremost Austronesian settlement in the Remote Oceania attested as early as 300 BC, the date we placed the incipience of the Tahitian language. Accordingly to many archaeological reconstructions[38,39,40], descendants from West Polynesia had spread through East Polynesian archipelagos and settled in Hawaii by 600 AD and in New Zealand by 1,000 AD testifying the earliest outset dates for the related languages. It is worth mentioning that all stride times between the offsets of these three Polynesian languages hold consistently the same rate of random phonetic changes $\alpha = (4.27 \pm 0.01) \cdot 10^{-4}$ affirming the validity of the 'adiabatic' conjecture described above.

The language divergence among Atayal people distributed throughout an area of rich topographical complexity is neatly organized by the myths of origin place, consanguine clans, and geographical barriers that have lead to the formation of a unique concept of ethnicity remarkable for such a geographically small region as Taiwan. The complexity of the Atayal ethnic system and the difficulty of defining the ethnic borders hindered the classification of the Atayal regional groups and their dialects which has been continuously modified throughout the last century. In our work, we follow the traditional classification[41] of the Atayal group into three branches based on their places of origin: Sediq (Sedek), Ciuli (Tseole) Atayal, and Squiliq (Sekilek) Atayal. In account with the standard lexicostatistic arguments, the Sediq dialect subgroup could split off from the rest of the Atayal groups about 1,600 years ago, as the both branches share up

to a half of the cognates in the 200 basic vocabulary wordlist[42]. This estimated date is very tentative in nature and calls for a thorough crosschecking. The Atayal people had been recognised as they had started to disperse to the northern part of Taiwan around 1,750 AD[43]. Being formed as the isolated dialect subgroups in island interiors, they showed the greatest diversity in race, culture, and social relations and sometimes considered each other as enemies and prime head hunting targets. Given the same rate of random phonetic changes α as derived for the Polynesian languages, the 'adiabatic' model of language evolution returns the stride times of 1,000 years between the Sediq dialect subgroup and Squiliq Atayal and of 860 years between the Ciuli and Squiliq Atayal languages. Consistently, Sediq is estimated to have branched off from the other Atayal languages 140 years before the main Atayal group split into two. The Squiliq subgroup had been attested during the latest migration of Atayal people, as late as 1,820 AD. Perhaps, a comprehensive study of the Atayal dialects by their symmetry can shed light on the origins of the Atayal ethnic system and its history.

We have presented the new network paradigm for the language phylogeny based on the analysis of language family symmetries that allows making accurate inferences on the most significant events of human history by tracking changes in language families through time. Geometric representations of language phylogenetic networks can be used in order to test the various statistical hypotheses about the evolution of languages. The simplest 'adiabatic' model can be refined by incorporating more fitting parameters, as the evolutionary mechanism is clarified. For instance, two words of different languages may become occasionally more similar by chance through a random replacement of phonemes. Although such an extraordinary event is obviously rare, with a very small probability of occurrence, its statistical impact may be remarkable especially over small vocabularies providing corrections for the derived chronological estimates. Computational simplicity of the proposed method based primarily on linear algebra (see the Method section) is its crucial advantage over previous approaches to the

computational linguistic phylogeny that makes it an invaluable tool for the automatic analysis of both the languages and the large document data sets that helps to infer on relations between them in the context of human history.

Methods Summary

The input to our analysis is the lexical distance matrix *D* constructed on the base of basic meanings collected by M. Swadesh[44] in the 1950s and used since then in lexicostatistics and glottochronology to evaluate the distances between pairs of languages. The symmetric matrix *D* uniquely determines a weighted undirected fully connected graph, in which vertices represent languages and edges connecting them have weights equal the lexical distances. To resolve its symmetry, we investigate linear operators defined by stochastic matrices invariant with respect to the graph automorphisms which are naturally interpreted as random walks preserving full information about the relations between languages embodied into the matrix of lexical distances. Random walks appeared in our analysis in concern to neither particular assumptions regarding to any evolutionary process, nor a Bayesian analysis previously used in linguistic phylogenetics[2,7,8], but as the unique linear transformation consistent with lexical symmetry of a language family. Random walks attribute a density function to each language convergent to a unique stationary distribution which does not correspond to any extant language, but defines a centre of symmetry of the sample described by *D*. In its turn, a density function representing a language may be rebuilt by a number of affine transformations originated from the centre of symmetry and weighted with probability to follow along a self-avoiding random walk path. The expected length of such a path called first-passage time[45] constitutes a natural structural distance on graphs that obeys the Euclidean relationships. Expected paths toward languages that cast in the same mould are coherent as revealed by geometric proximity

in Euclidean space spanned by the diffusion eigenmodes that might be either exploited visually, or accounted analytically.

Methods

Distance between languages by lexicostatistics

Contrary to traditional glottochronology where the percentage of shared cognates (words inherited from a common ancestor) in Swadesh vocabulary was suggested as an estimate of the closeness of two languages, we represent lexical substitutions by distances measured by sound changes, phoneme by phoneme [9,10]. The standard Levenshtein distance accounting for the minimal number of insertions, deletions, or substitutions of single letters needed to transform one word into the other used previously in information theory[46] gave questionable results while being applied to the automatic clustering of languages[47], since lengthy words provide more room for editing being therefore responsible for a decisive statistical impact distorting the results on language classification. In order to compare two words having the same meaning albeit different lengths, the actual edit distance has to be normalized by the number of characters of the longer of the two. For instance, the normalized Levenshtein distance between the German word *milch* and the English word *milk* equals 2/5. Then the lexical distance $D_{ij} \in [0,1]$ between two languages l_i and l_j have been computed as an average of normalized Levenshtein distances over Swadesh vocabulary of 200 meanings – the smaller the value is, the more affine are the languages.

(The symmetric matrices of lexical distances for the both language families are given in the **Supplementary Information.)**

Recovering symmetries of distance matrices by random walks

If two languages l_i and l_j are similar from a network perspective, the lexical distance between each of them and any other language in the sample is proximate, $D_{iv} \approx D_{jv}$, and remains proximate under any linear transformation of the matrix D compatible with its symmetry. Given all languages in the family equally similar, the set of automorphisms of the matrix D is the symmetric group of all permutations of the languages in the sample. However, the description of automorphisms in terms of group theory is cumbersome, so that we use another approach based on the investigation of random walks defined on the weighted undirected fully connected graph uniquely determined by the matrix D. In the simplest case, a random walk on D is defined by $T = \Delta^{-1} D$ where the diagonal matrix $\Delta = \mathrm{diag}\left(\sum_{v=1}^{50} D_{1,v}, \ldots \sum_{v=1}^{50} D_{50,v} \right)$ contains the cumulative lexical distances for each language. The matrix T is nothing else but a normalized matrix of lexical distances respecting the structure of graph and attributing a density function, $\sigma_i \geq 0, \sum_{i=1}^{50} \sigma_i = 1,$ to each language. Under the consequent actions of T, any density function σ converges to a stationary distribution $\pi = \lim_{n \to \infty} \sigma T^n$, but a fro way also exists. The eigenvectors $\{\psi_1, ..., \psi_{50}\}$ of the symmetric operator $\hat{T} = \pi^{1/2} T \pi^{-1/2}$ form an orthonormal basis in Hilbert space ordered with respect to the correspondent eigenvalues $1 = \mu_1 > \mu_2 \geq ... \geq \mu_{50} \geq -1$. The diffusion of random walkers over the graph is described by the self-adjoint Laplace operator $\hat{L} = 1 - \hat{T}$ which is invertible in the orthogonal complement of the Perron eigenvector $\psi_1 \hat{T} = \psi_1$ belonging to the largest eigenvalue $\mu_1 = 1$, so that $\hat{L}^{-1} = \sum_{k=2}^{50} \psi_{k,i} \psi_{k,j} / \psi_{1,i} \psi_{1,j} (1 - \mu_k)$ where $\psi_{k,i}$ is the i^{th} component of the eigenvector ψ_k, and all $\psi_{1,i} > 0$. The elements of the real symmetric matrix $\hat{L}^{-1}$ equal the expected lengths of the overlaps between random self-avoiding paths starting from a randomly chosen vertex of the graph, with the first-passage times on the diagonal. The matrix $\hat{L}^{-1}$ can be diagonalized by a real orthogonal matrix Q such that $\Lambda = Q^{\dagger} \hat{L}^{-1} Q$ is a diagonal matrix. Each column vector q_k of the matrix Q is an eigenvector of the linear transformation that determines a direction where $\hat{L}^{-1}$ acts as a simple rescaling, $\hat{L}^{-1} q_k = \theta_k q_k$ with some real eigenvalue $\theta_k \geq 0$. Contours of coherence of

random paths constitute ellipsoids centered at the stationary distribution, and eigenvectors q_k define the axes of these ellipsoids. In the diagrams shown in Fig.1a and Fig.3a, we have used the three major eigenvectors belonging to the three largest eigenvalues of the matrix $\hat{L}^{-1}$ as the Cartesian coordinates. Each density function representing a language was described by a point; the clusters of points formed jets, in which points corresponding to similar languages were geometrically proximate. In spherical coordinates, each language was described by three coordinates: a radius and two angles. If the two points representing two different languages share the same angle, there is a similarity transformation compatible with symmetry of the lexical distance matrix D that maps one point into another. In particular, from one hand, the Indo-Iranian, Balto-Slavic, Armenian, and Albanian languages sharing approximately the same zenith angle in the diagram shown in Fig.1a are similar, since there are the rotations of the IE caltrop in a plane around its azimuth axis that maps these spines one into another. From another hand, the Greek, Romance, Celtic, and Germanic languages are also similar as belonging to approximately the same azimuthal angle forming an isogloss which obviously coincides with the well-known Centum-Satem division of the Indo-European language family.

Figure Legends

Figure 1: How to draw a Fylfot? **a**, The sample of fifty IE languages in space of three major symmetry patterns colour coded. The figure origin is the common centre of symmetry, not the Proto-Indo-European language. The panels **b-c** show the kernel density estimates (in red) of the distributions of azimuthal and zenith angles (in radians) over the sample languages. Frequency histograms are given in blue. Values in plots indicate the absolute data frequencies. The grouping of the languages according to the Centum-Satem isogloss is indicated below the plots.

Figure 2: Tracing back the diverging pathways of language evolution. The panels **a-d.** show the normal probability plots fitting the distances from the common centre of symmetry to univariate normality. The values of variance are given for each language group. The coloured balls indicate the expected locations of protolanguages. **e,** The reconstruction of the Proto-Indo-European language breakup in space of three major symmetry patterns of the sample. **f,** The graphical test for checking the goodness of fit of the scaled distances from the proto-centre to multivariate normality.

Figure 3: Two independent branches of the Austronesian family diverge from Batanes. **a,** Fifty AU languages on the colour coded symmetry plane of the entire family. **b,** The kernel density estimate (in red) of the distribution of azimuthal angles (in radians) and the frequency histogram (in blue). **c,** Batanes, the islands where the AU family forks. **d,** Map of Austronesian expansion. Arrows mark the proposed routes. Language colours are the same as in **a.**

Supplementary Information accompanies the paper on **www.nature.com/nature**.

These authors contributed equally to this work.

Correspondence should be addressed to D.V. (volchenk@physik.uni-bielefeld.de).

---

[1] Nichols J., Warnow T. Tutorial on Computational Linguistic Phylogeny, *Language and Linguistics Compass* **2/5**, 760-820 (2008).

[2] Gray R.D, Atkinson Q.D. Language-tree divergence times support the Anatolian theory of Indo-European origin, Nature **426,** 435-439 (2003).

[3] Erdem E., Lifschitz V., Nakhleh L.& Ringe D. Reconstructing the evolutionary history of Indo-European languages using answer set programming, in *Proc. Practical Aspects of Declarative Languages (PADL)*, 160-176 (2003).

[4] Nakhleh L., Ringe D.& Warnow T. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages, *Language* **81**(2), 382-420 (2005).

[5] Blanchard Ph., Volchenkov D. *Mathematical Analysis of Urban Spatial Networks,* Springer Verlag (2008).

[6] Atkinson Q., Nichols G., Welch D.& Gray R. From words to dates: water into wine, mathematics, or phylogenetic inference. *Transactions of the Philological Society* **103,** 193-219 (2005).

[7] Gray R.D., Jordan F.M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052-1055 (2000).

[8] Gray R.D., Drummond A.J., & Greenhill S.J. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* **323**, 479-483 (2009).

[9] Petroni F., Serva M. Language distance and tree reconstruction, *J. Stat. Mech*. P08012 (2008).

[10] Serva M., Petroni F. Indo-European Languages tree by Levenshtein distance. *Europhys. Letters* **81**, 68005-9 (2008).

[11] Holman E.W. *et al*. Explorations in automated lexicostatistics. Folia Linguistica **42**(2), 331-354 (2008).

[12] Solta G.R. Palatalisierung und Labialisierung, *Indogermanische Forschungen* **70**, 276–315 (1965).

[13] Gamkrelidze T.V., Ivanov V.V. The early history of Indo-European language, *Scientific American* **262**(3), 110-116 (1990).

[14] Gimbutas M., *The Goddesses and Gods of Old Europe 7000 - 3500 BC, Mythos, Legends and Cult Images*, Thames & Hudson Ltd (1974).

[15] Chapman J. The Vinča culture of South-east Europe: studies in chronology, economy and society. *Oxford: British archaeological reports*, BAR international series **117** (1981).

[16] Schliemann H. *Troy and its remains*, London: Murray (1875).

[17] Greg R.P. Meaning and Origin of Fylfot and Swastika, *Archaeologia*, Vol. **XLVIII**, part 2 (1885).

[18] Mallory J.P. *In Search of the Indo-Europeans: Language, Archaeology and Myth*, Thames and Hudson (1989).

[19] Fouracre P. *The New Cambridge Medieval History*. Cambridge University Press, (1995-2007).

[20] Bryant E. *The Quest for the Origins of Vedic Culture: The Indo-Aryan Migration Debate*, Oxford University Press (2001).

[21] Novotná P., Blažek V. Glottochronology and its application to the Balto-Slavic languages. *Baltistica* **XLII** (2), 185–210 (2007).

[22] Mcleod J. *The History of India*, Greenwood Pub. Group (2002).

[23] Green P. *The Greco-Persian Wars*. Berkeley, Los Angeles, London: University of California Press (1996).

[24] Dahl O.C. Malgache et Maanjan: Une comparaison linguistique. *Avhandlinger utgitt av Egede-Instituttet*, vol. **3**, 408, Arne Gimnes Forlag (1951).

[25] Hurles M. E. *et al*. The Dual Origin of the Malagasy in Island Southeast Asia and East Africa: Evidence from Maternal and Paternal Lineages. *American Journal of Human Genetics* **76**, 894-901 (2005).

[26] Diamond J.M. Express train to Polynesia, *Nature* **336**, 307-308 (1988).

[27] Su. B. *et al.* Polynesian origins: Insights from the Y chromosome. *PNAS* **97** (15), 8225-8228 (2000).

[28] Bellwood P., Koon P. Lapita colonists leave boats unburned! The question of Lapita links with Island Southeast Asia. *Antiquity* **63** (240), 613–622 (1989).

[29] Kirch P.V. *The Lapita Peoples: Ancestors of the Oceanic World.* Cambridge, Mass., Blackwell (1997).

[30] Matisoo-Smith E., Robins J.H. Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat, *PNAS* **101** (24), 9167-9172 (2004).

[31] Larson G. *et al.* Phylogeny and ancient DNA of Sus provide insights into Neolithic expansion in Island Southeast Asia and Oceania. *PNAS* **104** (12), 4834-4839 (2007).

[32] Kirch P.V. *On the road of the winds: an archaeological history of the Pacific Islands before European contact*. University of California Press, Berkley, CA (2000).

[33] Anderson A., Sinoto Y. New radiocarbon ages of colonization sites in East Polynesia. *Asian Perspect.* **41**, 242-257 (2002).

[34] Hurles M.E. *et al*. Untangling Oceanic settlement: the edge of the knowable. *Trends Ecol. Evol.* **18**, 531-540 (2003).

[35] Lum J.K., Jorde L.B., Schiefenhovel W. Affinities among Melanesians, Micronesians, and Polynesians: a neutral biparental genetic perspective. *Hum. Biol.* **74**, 413–430 (2002).

[36] Kayser M. *et al.* Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol. Biol. Evol.* **23,** 2234–2244 (2006).

[37] Friedlaender J.S., *et al.* The genetic structure of Pacific Islanders. *PLoS Genet* **4**(1): e19. doi:10.1371/journal.pgen.0040019 (2008).

[38] Kirch P.V. *On the road of the winds: an archaeological history of the Pacific Islands before European contact*. University of California Press, Berkley, CA (2000).

[39] Anderson A., Sinoto Y. New radiocarbon ages of colonization sites in East Polynesia. *Asian Perspect.* **41**, 242-257 (2002).

[40] Hurles M.E. *et al*. Untangling Oceanic settlement: the edge of the knowable. *Trends Ecol. Evol.* **18**, 531-540 (2003).

[41] Utsurikawa N. *A Genealogical and Classificatory Study of the Formosan Native Tribes.* Tokyo: Toko shoin (1935).

[42] Li, P. J.-K. Types of lexical derivation of men's speech in Mayrinax. *BIHP* 54.3:1-18 (1983).

[43] Li, P. J.-K. The Dispersal of the Formosan Aborigines in Taiwan. *Language and Linguistics* **2**(1), 271-278 (2001).

[44] Swadesh M. Lexicostatistic dating of prehistoric ethnic contacts, *Proc. Am. Phil. Soc.* **96**, 452 (1952).

[45] Lovász L. Random Walks on Graphs: A Survey, *Bolyai Society Mathematical Studies* **2**, 1-46, Keszthely (1993).

[46] Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady* **10,** 707–710 (1966).

[47]Batagelj V., Pisanski T.& Keržič D. Automatic clustering of languages, *Comput. Linguistics* **18**(3) 339-352 (1992).